Σχολή Μηχανικών
Τμήμα Ηλεκτρολόγων και Ηλεκτρονικών Μηχανικών
Μάθημα: Υπολογιστική Νοημοσύνη - Εργαστήριο
Διδάσκοντες: Δρ. Αλεξανδρίδης Αλέξανδρος, Καθηγητής
Δρ. Φαμέλης Ιωάννης, Καθηγητής

3η Σειρά Ασκήσεων

(Παράδοση: Ομάδα Α μέχρι 27 Ιανουαρίου, Ομάδα Β μέχρι 3 Φεβρουαρίου)

"Will there come a time," said Man, "when data will be sufficient or is the problem insoluble in all conceivable circumstances?"
The Cosmic AC said, "NO PROBLEM IS INSOLUBLE IN ALL CONCEIVABLE CIRCUMSTANCES."
Man said, "When will you have enough data to answer the question?"
"THERE IS AS YET INSUFFICIENT DATA FOR A MEANINGFUL ANSWER."

Isaac Asimov, "The Last Question"

1η Άσκηση –Πρόβλεψη απόδοσης CPU – revisited

Το αρχείο Machine_CPU_NN.xlsx περιέχει δεδομένα με χαρακτηριστικά επεξεργαστών και την σχετική απόδοση τους, όπως αυτή υπολογίζεται από benchmark tests. Στόχος της άσκησης είναι να κατασκευαστεί ένα μοντέλο βασισμένο σε νευρωνικά δίκτυα MLP που να μπορεί να προβλέπει την σχετική απόδοση ενός επεξεργαστή με βάση τα χαρακτηριστικά του συστήματος. Συγκεκριμένα το μοντέλο θα είναι της μορφής:

$\hat{y} = NN(x_1, x_2, \ldots$, όπου:

NN: Η μη γραμμική συνάρτηση που υλοποιείται από το νευρωνικό δίκτυο

$\hat{y}$: Η πρόβλεψη του μοντέλου για τη σχετική απόδοση του επεξεργαστή

x_1: Χρόνος κύκλου (ns)

x_2: Ελάχιστη κεντρική μνήμη (KB)

x_3: Μέγιστη κεντρική μνήμη (KB)

x_4: Μνήμη Cache (KB)

x_5: Ελάχιστος αριθμός καναλιών

x_6: Μέγιστος αριθμός καναλιών

Α) Εκπαιδεύστε νευρωνικό δίκτυο το οποίο θα μαθαίνει τη συσχέτιση ανάμεσα στα χαρακτηριστικά του επεξεργαστή και τη σχετική του απόδοση. Θα πρέπει να χρησιμοποιήσετε τον παρακάτω χωρισμό των δεδομένων σε σύνολα:

- **Εκπαίδευση** Δεδομένα 1 μέχρι 90
- **Αξιολόγηση** Δεδομένα 91 μέχρι 150
- **Έλεγχος** Δεδομένα 151 μέχρι 209

Το νευρωνικό δίκτυο θα πρέπει να έχει 2 κρυφές στοιβάδες με 5 νευρώνες στην πρώτη και 3 νευρώνες στη δεύτερη στοιβάδα. Τα δεδομένα θα πρέπει να δίνονται στο νευρωνικό δίκτυο κανονικοποιημένα. Κατασκευάστε πρόγραμμα τύπου function το οποίο να παίρνει σαν εισόδους το σύνολο των διαθέσιμων δεδομένων και να επιστρέφει το εκπαιδευμένο δίκτυο, καθώς και τους στατιστικούς δείκτες MARE% και R^2 στα σύνολα αξιολόγησης και ελέγχου.

Για τον υπολογισμό των MARE% και R^2 χρησιμοποιήστε τους παρακάτω τύπους:

$$MARE\% = 100\frac{\sum_{j=1}^{P}\frac{|y_j - \hat{y}_j|}{y_j}}{P} \text{ και}$$

$$R^2 = 1 - \frac{SSE}{SST}, \text{ όπου } SSE = \sum_{j=1}^{P}\left(y_j - \hat{y}_j\right)^2 \text{ και } SST = \sum_{j=1}^{P}\left(y_j - \bar{y}\right)^2$$

ΠΡΟΣΟΧΗ: Πριν καλέσετε τη συνάρτηση feedforwardnet για να δημιουργήσετε το δίκτυο, θα πρέπει πρώτα να εκτελέσετε την εντολή:
rand('seed',10);

Η εντολή αυτή δίνει συγκεκριμένες τιμές στις τυχαίες επιλεγόμενες αρχικές τιμές των παραμέτρων του δικτύου, ώστε η λύση που τελικά θα βρεθεί να είναι ικανοποιητική.

Β) Εάν ο αριθμός νευρώνων στη πρώτη κρυφή στοιβάδα κυμαίνεται από 5 έως 20 και στη δεύτερη στοιβάδα από 3 έως 10, εκπαιδεύστε όλους τους πιθανούς συνδυασμούς νευρωνικών δικτύων που προκύπτουν αποθηκεύοντας κάθε φορά τους στατιστικούς δείκτες (Θα πρέπει και πάλι κάθε φορά που δημιουργείτε το δίκτυο να εκτελείτε την εντολή rand('seed',10);). Παρατηρείτε σημαντικές αποκλίσεις στις τιμές των στατιστικών δεικτών από εκτέλεση σε εκτέλεση; Εάν ναι, που πιστεύετε ότι οφείλεται αυτό;

Γ) Από όλα τα νευρωνικά δίκτυα που αποθηκεύσατε στο Β ερώτημα, επιλέξτε αυτό που έδωσε τον καλύτερο συντελεστή R^2 στο σύνολο δεδομένων ελέγχου. Στη συνέχεια, θεωρήστε ότι για κάποιον επεξεργαστή ισχύουν οι ακόλουθες τιμές:

Χρόνος κύκλου: 200 ns

Ελάχιστη κεντρική μνήμη: 3000 KB

Ελάχιστος αριθμός καναλιών: 6

Μέγιστος αριθμός καναλιών: 16

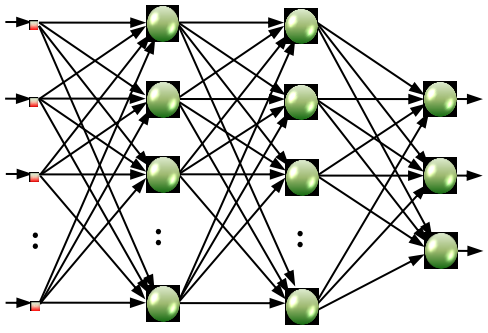
Κατασκευάστε στο Matlab τρισδιάστατη γραφική παράσταση που να απεικονίζει τις προβλέψεις του καλύτερου νευρωνικού δικτύου για την απόδοση του επεξεργαστή (άξονας Z), συναρτήσει της μέγιστης κεντρικής μνήμης (άξονας X) και της μνήμης Cache (άξονας Y). Θεωρήστε ότι μέγιστη κεντρική μνήμη μεταβάλλεται από τα 8000KB μέχρι τα 16000KB και η μνήμη Cache μεταβάλλεται από 32KB μέχρι 128KB.

Με βάση το γράφημα που κατασκευάσατε, σχολιάστε την επίδραση της κεντρικής μνήμης και της μνήμης Cache στην απόδοση του επεξεργαστή. Συγκρίνετε το γράφημα που κατασκευάστε με το αντίστοιχο γράφημα που είχατε κατασκευάσει στη 2η σειρά ασκήσεων και σχολιάστε τις διαφορές τους.

Δ) Επεκτείνετε το τρισδιάστατο γράφημα για μεταβολή της μέγιστης κεντρικής μνήμης από 8000KB μέχρι 64000KB και της μνήμης Cache από 32KB μέχρι 512KB. Σχολιάστε το γράφημα. Θεωρείτε αξιόπιστες τις προβλέψεις του νευρωνικού δικτύου σε αυτή την περίπτωση;

Bonus material:

- www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html
- www.statsoft.com/textbook/neural-networks/
- www.learnartificialneuralnetworks.com/backpropagation.html
- el.wikipedia.org/wiki/Νευρωνικό_δίκτυο



2η Άσκηση – Αυτόματος ανιχνευτής spam emails

Στόχος της άσκησης είναι να κατασκευαστεί ένα φίλτρο που να αποφασίζει αυτόνομα πότε ένα email είναι spam και πότε όχι. Το φίλτρο θα βασίζεται σε νευρωνικό δίκτυο MLP το οποίο θα λαμβάνει σαν εισόδους διάφορα στοιχεία σχετικά με τους χαρακτήρες και τις λέξεις που υπάρχουν μέσα στο email (π.χ. συχνότητα εμφάνισης συγκεκριμένων λέξεων, συχνότητα εμφάνισης συγκεκριμένων χαρακτήρων, κτλ.) και θα επιστρέφει την απόφαση σχετικά με το αν το email είναι spam ή όχι.

Τα διαθέσιμα δεδομένα[1] για την κατασκευή του δικτύου βρίσκονται στο αρχείο "spamdata.mat". Οι πρώτες 57 στήλες του πίνακα dat περιέχουν τις μεταβλητές εισόδου του δικτύου ενώ η τελευταία στήλη περιέχει την έξοδο που μπορεί να λαμβάνει δύο τιμές:

- Τιμή 1: Το συγκεκριμένο email είναι spam
- Τιμή 0: Το συγκεκριμένο email δεν είναι spam

Συνολικά είναι διαθέσιμα 4601 δεδομένα.

Α) Εκπαιδεύστε νευρωνικό δίκτυο με δύο κρυφές στοιβάδες (20 νευρώνες στην πρώτη κρυφή στοιβάδα και 10 νευρώνες στη δεύτερη) το οποίο θα λαμβάνει ως εισόδους τις 57 μεταβλητές εισόδου του προβλήματος και θα επιστρέφει την πρόβλεψη σχετικά με το αν ένα email είναι spam ή όχι. Τα δεδομένα θα πρέπει να αναδιαταχθούν με τυχαίο τρόπο και στη συνέχεια για την εκπαίδευση του δικτύου θα πρέπει να χρησιμοποιηθούν τρία υποσύνολα:

- Υποσύνολο εκπαίδευσης (40% των δεδομένων)
- Υποσύνολο αξιολόγησης (30% των δεδομένων)
- Υποσύνολο ελέγχου (30% των δεδομένων)

Τα δεδομένα εισόδου θα πρέπει να δίνονται στο νευρωνικό δίκτυο κανονικοποιημένα. Κατασκευάστε πρόγραμμα τύπου function το οποίο να παίρνει σαν εισόδους το σύνολο των διαθέσιμων δεδομένων και να επιστρέφει το εκπαιδευμένο δίκτυο, καθώς και τα αποτελέσματα στην εξής μορφή:

[1]Τα δεδομένα έχουν ληφθεί από το UCI Machine Learning Repository, όπου και υπάρχουν διαθέσιμες περισσότερες πληροφορίες σχετικά με το συγκεκριμένο dataset και τις μεταβλητές εισόδου που χρησιμοποιούνται: *http://archive.ics.uci.edu/ml/datasets/Spambase*

- Συνολικό ποσοστό emails που κατηγοριοποιήθηκαν σωστά σε οποιαδήποτε από τις δύο κατηγορίες, ξεχωριστά για τα σύνολα αξιολόγησης και ελέγχου
- Ποσοστό emails που είναι spam και κατηγοριοποιήθηκαν σωστά ως spam (Spam-Correct) ξεχωριστά για τα σύνολα αξιολόγησης και ελέγχου
- Ποσοστό emails που δεν είναι spam και κατηγοριοποιήθηκαν σωστά ως μη spam (Not Spam-Correct) ξεχωριστά για τα στα σύνολα αξιολόγησης και ελέγχου
- Ποσοστό emails που είναι spam και κατηγοριοποιήθηκαν λάθος ως μη spam (Spam Detection Failure) ξεχωριστά για τα στα σύνολα αξιολόγησης και ελέγχου
- Ποσοστό emails που δεν είναι spam και κατηγοριοποιήθηκαν λάθος ως spam (Spam False Alarm) ξεχωριστά για τα στα σύνολα αξιολόγησης και ελέγχου

Παρουσιάστε τον πίνακα σύγχυσης (confusion matrix) για τα τρία σύνολα δεδομένων.

ΠΡΟΣΟΧΗ: Τα ανωτέρω ποσοστά θα πρέπει να υπολογιστούν αναλυτικά μέσα στο function που θα κατασκευάσετε και όχι να ληφθούν έτοιμα από το Matlab.

Β) Επαναλάβετε τη διαδικασία της εκπαίδευσης (κρατώντας σταθερό το διαχωρισμό στα τρία υποσύνολα). Αφού εκτελέσετε έναν αριθμό επαναλήψεων, επιλέξτε το μοντέλο που κρίνετε ότι ανταποκρίνεται καλύτερα στη συγκεκριμένη εφαρμογή και αιτιολογήστε την επιλογή σας.

Bonus material:

- archive.ics.uci.edu/ml/datasets/spambase
- prag.diee.unica.it/pra/eng/research/doccategorisation/spamfiltering/bibliography
- homepages.cae.wisc.edu/~ece539/project/f03/sivanadyan.pdf
- www.aueb.gr/users/ion/docs/michelakis_final_report.pdf